



Home



List

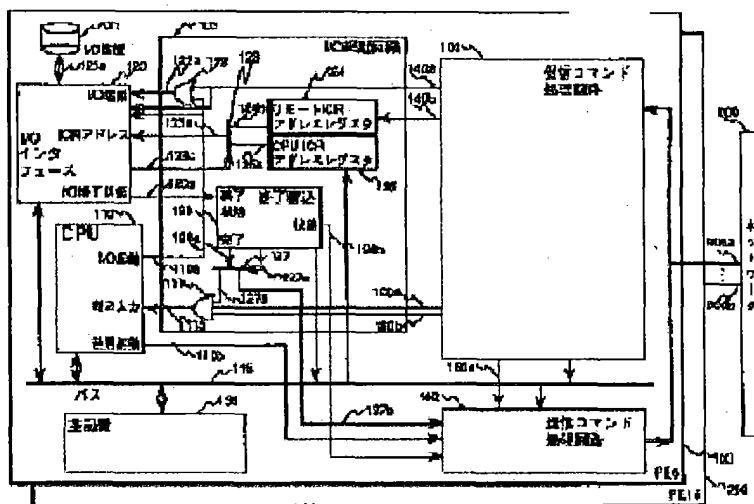
☐ Include

## MicroPatent® PatSearch FullText: Record 1 of 1

Search scope: JP ; Full patent spec.

Years: 1995-2003

Text: Patent/Publication No.: JP10021203



Order This Patent

Family Lookup

Find Similar

Legal Status

[Go to first matching text](#)

JP10021203 A

ACCESS METHOD FOR I/O DEVICE AND MULTIPROCESSOR SYSTEM FOR THE SAME  
HITACHI LTDInventor(s): TARUI TOSHIKI ; KITAI KATSUYOSHI ; MASHEL FREDERICO ; HIGUCHI TATSUO ; MURAHASHI HIDEKI  
Application No. 08176863 JP08176863 JP, Filed 19960705, A1 Published 19980123**Abstract: PROBLEM TO BE SOLVED:** To access an I/O device in another processing element(PE) without the intervention of the PE.**SOLUTION:** A file system of an access-source PE issues an I/O instruction specifying an I/O control code, a device driver sends a command requesting the execution of I/O operation that the instruction requests to an access-destination PE through a network 900 if an I/O device that the instruction requests is at another PE. At the access-destination PE, a receive command processing circuit 101 and an I/O actuating circuit 103 generates an I/O control record for performing the I/O operation at the request without the intervention of an OS and actuates an

I/O interface circuit 120. After this command is executed, a transmit command processing circuit 102 returns a return command corresponding to the command.

Int'l Class: G06F015163;

Patents Citing this One: No US, EP, or WO patents/search reports have cited this patent.



Home



List

---

For further information, please contact:  
[Technical Support](#) | [Billing](#) | [Sales](#) | [General Information](#)

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-21203

(43) 公開日 平成10年(1998) 1月23日

(51) Int.Cl.<sup>6</sup>

識別記号

庁内整理番号

F I

技術表示箇所

G 0 6 F 15/163

G 0 6 F 15/16

3 2 0 S

審査請求 未請求 請求項の数14 O L (全 17 頁)

(21) 出願番号 特願平8-176863

(22) 出願日 平成 8 年(1996) 7 月 5 日

(71) 出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目 6 番地

(72) 発明者 垂井 俊明

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 北井 克佳

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(72) 発明者 マシエル・フレデリコ

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(74) 代理人 弁理士 高橋 明夫

最終頁に続く

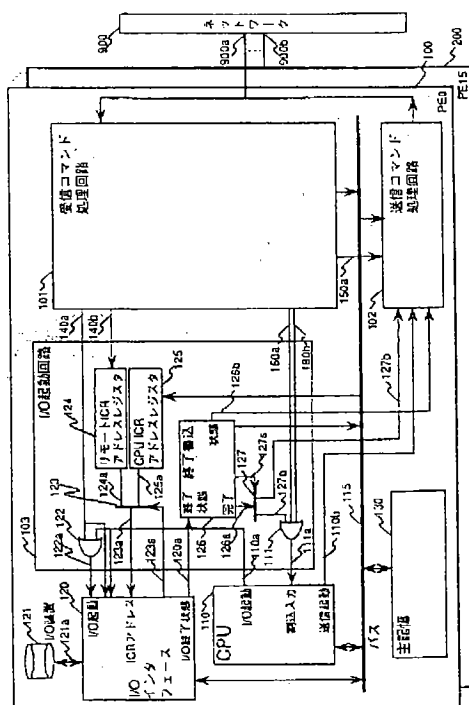
(54) 【発明の名称】 I/O装置のアクセス方法およびそのためのマルチプロセッサシステム

(57) 【要約】 (修正有)

【課題】 他のPEのOSの介入を経ないでそのPE内のI/O装置をアクセスする。

【解決手段】 アクセス元PEのファイルシステムが、I/Oコントロールレコードを指定するI/O命令を発行し、デバイスドライバは、その命令が要求するI/O装置が他のPEにあるときには、その命令が要求するI/O動作の実行を要求するコマンドをネットワーク900を介してアクセス先のPEに送信する。アクセス先のPEでは、受信コマンド処理回路101とI/O起動回路103とが、この要求にตอบสนองしてこのI/O動作を実行するためのI/OコントロールレコードをOSの介入を得ないで作成し、I/Oインタフェース回路120を起動する。このコマンドの実行後に、送信コマンド処理回路102はこのコマンドに対する返信コマンドを返送する。

(図1a)



## 【特許請求の範囲】

【請求項1】第1、第2のプロセッシングエレメントと、該第1、第2のプロセッシングエレメントを結ぶデータ転送用のネットワークを有し、

該第1、第2のプロセッシングエレメントの各々は、主記憶と、少なくとも一つの処理装置と、少なくとも一つのI/O装置とを有するマルチプロセッサシステムにおいて、

上記第1のプロセッシングエレメントのOSが、上記第2のプロセッシングエレメント内のI/O装置へのアクセスを要求する前に、該第1のプロセッシングエレメントのそのOSが、該第2のプロセッシングエレメントのOSとの間でメッセージを上記ネットワークを介して交換することにより、上記第2のプロセッシングエレメント内の上記I/O装置のアクセス許可を取得し、

そのアクセス許可の取得後に上記第1のプロセッシングエレメントのOSが、上記第2のプロセッシングエレメント内のI/O装置へのアクセスを要求し、

該第1のプロセッシングエレメントのOSによる上記要求に応答して、該第2のプロセッシングエレメント内の上記I/O装置におけるI/O動作を要求するコマンドを上記第1のプロセッシングエレメントから上記第2のプロセッシングエレメントに上記ネットワークを介して転送し、

上記第2のプロセッシングエレメントにおいて、上記コマンドに応答して、その第2のプロセッシングエレメント内の上記I/O装置に上記コマンドにより要求された上記I/O動作の実行を、該第2のプロセッシングエレメント内の上記OSを介さないで直接指示するI/O装置のアクセス方法。

【請求項2】上記第2のプロセッシングエレメントにおいて、上記I/O動作の実行結果を表す返信用コマンドを、該第2のプロセッシングエレメント内のOSの介入を得ないで作成し、

作成されたコマンドを上記第2のプロセッシングエレメントから該第1のプロセッシングエレメントに上記ネットワークを介して転送するステップをさらに有する請求項1記載のI/O装置のアクセス方法。

【請求項3】第1、第2のプロセッシングエレメントと、該第1、第2のプロセッシングエレメントを結ぶデータ転送用のネットワークを有し、  
該第1、第2のプロセッシングエレメントの各々は、主記憶と、少なくとも一つの処理装置と、少なくとも一つのI/O装置と、そのプロセッシングエレメントを制御するOSから発行された、該I/O装置へのアクセス要求に応答して、該アクセス要求が指定するI/O動作指定情報に基づいて、該I/O装置をアクセスするI/O制御装置とを有するマルチプロセッサシステムにおいて、

上記第1のプロセッシングエレメント内のOSから発行

された、上記第2のプロセッシングエレメント内のI/O装置へのアクセス要求に応答して、該I/O動作の実行を要求するコマンドを上記第1のプロセッシングエレメントから上記第2のプロセッシングエレメントに上記ネットワークを介して転送し、

上記第2のプロセッシングエレメントにおいて、上記コマンドに応答して、その第2のプロセッシングエレメント内の上記I/O装置に上記I/O動作の実行を指示するためのI/O動作指定情報を、該第2のプロセッシングエレメント内の上記OSを介さないで直接生成し、生成されたI/O動作指定情報を用いて該第2のプロセッシングエレメント内の該I/O制御装置を上記OSを介さないで直接起動するI/O装置のアクセス方法。

【請求項4】上記第2のプロセッシングエレメントにおいて、上記I/O動作の実行結果を表す返信用コマンドを、該第2のプロセッシングエレメント内のOSの介入を得ないで作成し、

作成されたコマンドを上記第2のプロセッシングエレメントから該第1のプロセッシングエレメントに上記ネットワークを介して転送するステップをさらに有する請求項3記載のI/O装置のアクセス方法。

【請求項5】第1、第2のプロセッシングエレメントと、

該第1、第2のプロセッシングエレメントを結ぶデータ転送用のネットワークを有し、

該第1、第2のプロセッシングエレメントの各々は、主記憶と、

少なくとも一つの処理装置と、

少なくとも一つのI/O装置と、

そのプロセッシングエレメントを制御するOSから発行された、該I/O装置へのアクセス要求に応答して、該アクセス要求が指定するI/O動作指定情報に基づいて、該I/O装置をアクセスするI/O制御装置とを有し、

該第1のプロセッシングエレメントは、

該第1のプロセッシングエレメントのOSから発行された、該第2のプロセッシングエレメント内の該I/O装置に対するアクセス要求に応答して、そのアクセス要求が指定するI/O動作の実行を該第2のプロセッシングエレメントに要求する少なくとも一つのコマンドを生成するコマンド生成回路と、

該コマンドを該第2のプロセッシングエレメントに該ネットワークを介して送信する回路とをさらに有し、

該第2のプロセッシングエレメントは、該第1のプロセッシングエレメントから上記コマンドが転送されたときに、該第2のプロセッシングエレメント内の上記I/O装置に上記コマンドが要求するI/O動作を実行させるためのI/O動作指定情報を、該コマンドに基づいて、かつ、該OSの介入を得ないで生成し、該生成されたI/O動作指定情報を用いて該第2のプロセッシングエレ

メント内の上記I/O制御装置を起動するI/O動作指定情報作成回路をさらに有するマルチプロセッサシステム。

【請求項6】該第2のプロセッシングエレメントは、上記作成されたI/O動作指定情報に基づいて該第2のプロセッシングエレメント内の該I/O制御装置が、該第2のプロセッシングエレメント内の該I/O装置におけるI/O動作が完了した時点で、該I/O動作の実行結果を該第1のプロセッシングエレメントに通知するための返信用コマンドを作成する返信用コマンド作成回路と、

該作成された返信用コマンドを該第1のプロセッシングエレメントに該ネットワークを介して送信する回路とをさらに有する請求項5記載のマルチプロセッサシステム。

【請求項7】上記返信用コマンドは、上記第1のプロセッシングエレメントから転送された上記コマンドが書き込み動作を要求するとき、その要求された書き込み動作が正常に終了したか否かの終了状態情報を含み、上記第1のプロセッシングエレメントから転送された上記コマンドが読み出し動作を要求するとき、その要求された読み出し動作が正常に終了したか否かの終了状態情報とその要求された読み出し動作が正常に終了したときにはそのコマンドで要求されたデータとを含む請求項6記載のマルチプロセッサシステム。

【請求項8】上記コマンドは、上記第1のプロセッシングエレメントのOSが発行した上記アクセス要求が指定したI/O動作指定情報を含み、上記I/O動作指定情報作成回路は、上記転送されたコマンドに含まれた上記I/O動作指定情報を用いて、上記コマンドが要求するI/O動作を実行させるためのI/O動作指定情報を作成する回路を有する請求項5記載のマルチプロセッサシステム。

【請求項9】上記I/O動作指定情報作成回路により生成される上記I/O動作指定情報は、該第2のプロセッシングエレメント内のOSが発行するI/O装置へのアクセス要求が指定するI/O動作指定情報と同じ種類の複数の項目を含む請求項8記載のマルチプロセッサシステム。

【請求項10】該第1、第2のプロセッシングエレメントの各々内のOSは、ファイルシステムとデバイスドライバを含み、該ファイルシステムがファイルシステムレベルのI/O命令を発行し、該デバイスドライバは、このファイルシステムレベルのI/O命令が指定するI/O装置が、そのプロセッシングエレメント内にあるときには、このI/O命令が要求するI/O動作をそのプロセッシングエレメント内の上記I/O制御装置に指示するための、デバイスドライバレベルのI/O命令を発行するように構成され、さらにこのデバイスドライバは、上記ファイルシステムレベルのI/O命令が指定するI

/O装置が、他のプロセッシングエレメント内にあるときには、このI/O命令が要求するI/O動作を指定するI/O動作指定情報を指定し、そのI/O動作の実行を当該他のプロセッシングエレメント内の上記I/O制御装置に要求するためのアクセス要求を発行するように構成され、

上記I/O動作指定情報作成回路により生成される上記I/O動作指定情報は、該第2のプロセッシングエレメント内のOSのデバイスドライバレベルのI/O命令が指定するI/O動作指定情報と同じ種類の複数の項目を含み、かつ、該デバイスドライバレベルのI/O命令が指定するI/O動作指定情報と同じデータ構造を有する請求項8記載のマルチプロセッサシステム。

【請求項11】該第1のプロセッシングエレメント内のOSが発行するアクセス要求が指定するI/O動作指定情報は、データの読み出しとデータの書き込みの一方を指定する動作種別情報と、アクセスすべきI/O装置内のアドレスを指定するアクセス位置情報と、書き込むべきデータあるいは読み出されたデータを格納すべき、バッファ領域のアドレスを指定するバッファアドレス情報とを含み、

該生成されるコマンドは、上記アクセス要求が指定する上記I/O動作指定情報の内の上記動作種別指定情報と、上記アクセス位置情報を少なくとも含み、該第2のプロセッシングエレメントは、上記第1のプロセッシングエレメントから要求されたI/O動作に使用する書き込みデータあるいは読み出しデータを一時的に格納するためのバッファ領域をさらに有し、

上記I/O動作指定情報作成回路により作成される上記I/O動作指定情報は、上記転送されたコマンドに含まれた上記動作種別指定情報と、上記アクセス位置情報と、上記バッファ領域のアドレスを含む請求項8記載のマルチプロセッサシステム。

【請求項12】該第1のプロセッシングエレメントは、上記I/O命令がデータの書き込みを要求するときに、上記コマンドとは別に上記I/O命令が要求する書き込みデータを含む他のコマンドを該第2のプロセッシングエレメントに転送する回路をさらに有し、該第2のプロセッシングエレメントは、該他のコマンドが転送されたときに該バッファ領域を確保し、そのバッファ領域に該転送された他のコマンド内の書き込みデータを転送する回路をさらに有し、該I/O動作指定情報作成回路は、該確保されたバッファ領域のアドレスをさらに使用して上記I/O動作指定情報を作成する回路を有する請求項11記載のマルチプロセッサシステム。

【請求項13】該第2のプロセッシングエレメントは、該転送されたコマンドがデータの読み出しを要求するときに、読み出すべきデータを保持する領域として上記バッファ領域を確保する回路をさらに有し、

該I/O動作指定情報作成回路は、該確保されたバッファ領域のアドレスをさらに使用して上記I/O動作指定情報を作成する回路を有する請求項1記載のマルチプロセッサシステム。

【請求項14】該第2のプロセッシングエレメントは、該データ読み出しのI/Oアクセスが終了した際に、該確保されたバッファ上に読み出されたデータを、該第1のプロセッサに返送するためのネットワークコマンドを作成する回路をさらに有する請求項13記載のマルチプロセッサシステム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明はI/O装置のアクセス方法およびそのためのマルチプロセッサシステムに関する。

【0002】

【従来の技術】計算機性能の飛躍的向上に関して、多数台のプロセッシングエレメントを並列動作させる、並列計算機が有望視されている。以下ではプロセッシングエレメントを略してPEと呼ぶ。とくに、独立した主記憶を持つPEをプロセッサ間ネットワークにより接続した疎結合並列計算機が、その台数拡張性により、注目されている。

【0003】ここで、疎結合の並列計算機においては、ディスクなどのI/O装置は各PEに分割して配置されるシェアードナッシングアーキテクチャが主流である。したがって、ある各PEがI/O装置への入出力アクセス（以下ではI/Oアクセスと呼ぶ）を行う場合、アクセスしようとするI/O装置が他のPEにおかれている場合がある。したがって、何らかの方法で他のPEにI/Oを依頼しなければならない。

【0004】I/Oの一般的な動作は以下の通りである。アプリケーションプログラムが、ファイルへのアクセスを発行すると、アクセス要求は、まずOSの一部であるファイルシステムに伝えられる。ファイルシステムでは、ファイルの論理名を物理的なファイルの位置に変換する。それと同時にファイルシステムは、論理的なファイルに対するロック操作などのファイルの論理的な整合性を保持するための処理を行う。ファイルシステムの処理が行われた後、アクセス要求は、OSの下位の機構である、デバイスドライバに伝えられる。デバイスドライバは、ファイルの物理的な記憶位置を入力として、実際に物理的なディスク装置をアクセスするためのプログラムである。ここで、デバイスドライバは、実際の装置をアクセスするために、物理的なI/O装置の機種に依存するのに対し、ファイルシステムの部分はI/O装置の機種に依存しない論理的な処理を行う。

【0005】従来の代表的なI/Oアーキテクチャは、「IBM: System 370 Principle of Operation」において述べられているチ

ヤネル方式がある。I/O装置へのアクセスは、CPUとは独立したI/O専用のプロセッサであるチャンネル

(I/Oインタフェース回路)により行われる。チャンネル方式においては、CPUからチャンネルへの命令の伝達は、物理的なファイルの位置を指定してアクセスが行わなければならない。従って、チャンネル装置はデバイスドライバにより起動される。CPUからチャンネルへの命令の伝達は以下の手順で行われる。プログラムがファイルにアクセスすると、ファイルシステムは、ファイルの物理的な記憶位置(データアドレス等)を求める。その後、デバイスドライバは、主記憶上にチャンネルコントロールワード(CCW)と呼ばれる構造体を作成し、I/Oアクセスの種類、データアドレス、データサイズ等の、チャンネルにI/Oアクセスを指示するための命令を書き込む。その後、デバイスドライバはI/Oアクセス命令を発行し、チャンネルをトリガする。チャンネルは、CCW上のアクセスパラメータを読み出し、その指示に基づいて、データの入出力を行う。アクセスが終了するとチャンネルはI/Oの終了状態を主記憶上の固定領域に出力するとともに、CPUに割込を掛け、I/Oの終了を知らせる。このチャンネルによる入出力方式は、CPUの処理と独立にI/Oアクセスを行うことができるため、広く用いられている。

【0006】現在のパソコン、ワークステーション等におけるI/Oアクセスの多くも、基本的にはチャンネルの方式を踏襲しており、以下の手順で行われる。

【0007】(イ)ファイルシステムが論理的なファイル名より、物理的なファイルの位置を求める。

【0008】(ロ)デバイスドライバが主記憶上にI/Oを指示するためのコントロールレコード(上記のCCWに相当する)を作成する。以下ではこのコントロールレコードをICR(I/O Control Record)と呼ぶ。

【0009】(ハ)デバイスドライバがI/Oインタフェース回路(SCSIインタフェース回路等、上記のチャンネルに相当する)をトリガする。

【0010】(ニ)I/Oインタフェース回路が主記憶上のコントロールレコードを読み出し、実際にI/O装置をアクセスする。

【0011】(ホ)デバイスドライバはI/O終了割り込みを受け、アクセス結果をファイルシステムを介してアプリケーションに戻す。

【0012】従来の疎結合の並列計算機において、他のPEの持つI/O装置に対してアクセスを行なおうとする場合、アクセス元のCPUで動作するデバイスドライバは、アクセス先のI/Oインタフェース回路に対して直接指示を与えることはできない。よって、I/O装置を持つPEに対してI/O装置をアクセスすることをデバイスドライバの間の、プロセッサ間通信を用いて依頼しなければならない。

【0013】したがって、以下の手順でアクセスを行う必要がある。

#### 【0014】A ファイルシステムの処理

(A1) アクセス元のPEのファイルシステムで、ファイルの論理的な名前より、物理的な記憶位置を求めようとした結果、ファイルは他のPEの持つディスク上に存在すると判断される。

【0015】(A2) アクセス元のPEのファイルシステムは、I/O装置を持つPEのファイルシステムとプロセッサ間通信を用いて相談を行い、ファイルの物理的な記憶位置を教えてもらうとともに、ファイルのアクセス権を獲得する。

【0016】(A3) アクセス元のPEのファイルシステムは、デバイスドライバに対して、(A2)で得られたファイルの位置などの情報を用いて、他のPEの持つファイルをアクセスすることを依頼する。

【0017】(A4) アクセス元のPEのデバイスドライバは、I/O装置を持つPEのデバイスドライバに対して、プロセッサ間通信を用いて、ファイルの物理的な記憶位置を指定して、ファイルをアクセスすることを依頼する。

#### 【0018】B I/O動作の実行

(B1) I/O装置を持つPEのOSは、処理(A4)のプロセッサ通信を受け取ると、メッセージ受信割込処理の中で、他のPEのデバイスドライバからI/Oアクセスを依頼されたことを検出する。そこで、依頼されたI/Oアクセスを処理するために、プロセッサ間通信で送られてきたファイルの物理的位置などの情報を用いて、デバイスドライバを起動する。

【0019】(B2) I/O装置を持つPEのデバイスドライバは、主記憶上にI/Oを指示するためのコントロールレコード(ICR)を作成する。

【0020】(B3) I/O装置を持つPEのデバイスドライバがI/Oインタフェース回路をトリガする。

【0021】(B4) I/O装置を持つPEのI/Oインタフェース回路が主記憶上のコントロールレコードを読み出し、実際にI/O装置をアクセスする。

【0022】(B5) I/Oアクセスが終了し、I/O終了割り込みを受けた、I/O装置を持つPEのデバイスドライバは、プロセッサ間通信を用いて、I/Oアクセス結果を、アクセス元のPEに送る。

#### 【0023】C I/O実行結果の取り込み

(C1) アクセス元のプロセッサのデバイスドライバは(B59)で送られた結果を、ファイルシステムを介して、アプリケーションに戻す。

【0024】以上の手順により、任意のPEが、任意のI/O装置をアクセスすることが可能になり、データの位置を意識しない柔軟なプログラミングが実現できる。但し、同じファイルを何回もアクセスした場合には、ファイルシステム管理のための処理(A1)(A2)は省

くことができ、実際にデータをアクセスするためのデバイスドライバ間の処理(A3)～(C1)のみを行えばよい。

#### 【0025】

【発明が解決しようとする課題】上記従来技術では、他のPEへのI/Oアクセスが行われた場合、I/O装置を持つPEのOSは、上記処理(B1)から(B3)および処理(B5C)という二つの中継処理を実行しなければならない。このため、このPEが本来実行していた仕事が滞るという問題点がある。

【0026】さらに、これらの処理はメッセージの到着割込、I/Oの終了割込という非同期的な割込を契機にOSが実行しなければならないため、上記の処理のオーバーヘッドの他に、割込によるアプリケーションプログラムからOSへの空間切替の非常に重いオーバーヘッドも発生する。

【0027】従来の科学技術用の並列計算機においては、プログラムの挙動は予め予想することができる、定型的な処理が中心である。したがって、上記の問題に対しては、I/O装置のデータをあらかじめアクセスを行うPEに割り振ることにより他のPEの持つI/O装置へのアクセスをなるべく抑えたり、他のPEのI/O装置へのアクセスをなるべく大きな単位でまとめて行う等の工夫を行い、他PEのI/O装置へのアクセス回数を極力おさえることが可能であった。

【0028】しかし、近年実用化されている、データベースなどのプログラムを並列マシンで実行する場合、I/O装置へのアクセスはデータベースの内容、データベースへの検索条件により大きく変化するため、予想することはできない。したがって、科学技術計算で行われていたような、データの分割、アクセスのブロック化などの手法で、他PEへのI/Oアクセスを削減することは不可能である。従って、他のPEの持つI/O装置へのアクセスが頻発してしまうと考えられる。特に、特定のPEへのI/Oアクセスが集中すると、アクセスされたPEのCPUがほとんど他PEのI/Oの中継処理のために使われてしまい、プログラムの実行効率が大幅に低下する。

【0029】従って、本発明の目的は、あるPEが他のPEの持つI/O装置をアクセスする際に、当該他のPEのOSの介入を得ないでそのI/O装置をアクセスできるマルチプロセッサシステムを提供することである。

#### 【0030】

【課題を解決するための手段】上記目的を達成するために、本発明では、アクセス先のPEのOSが行っていた中継処理を、そのOSの介入を得ないで直接ハードウェアで実行させる。

【0031】すなわち、本発明によるI/O装置アクセス方法では、アクセス元のPEのOSが、他のPE内のI/O装置へのアクセスを要求する前に、アクセス元の

OSからアクセス先のPE内のI/O装置をアクセスする許可を、アクセス元のOSが、アクセス先のOSとの間で上記ネットワークを介したメッセージパッシングにより取得する。

【0032】このアクセス許可の取得後に、アクセス先のPE内のI/O装置におけるI/O動作を要求するコマンドをアクセス元のPEからアクセス先のPEに、これらのPEを結ぶネットワークを介して転送する。アクセス先のPEにおいては、上記コマンドにตอบสนองして、そのPE内のI/O装置に上記コマンドにより要求されたI/O動作の実行をアクセス先のPE内のOSの介入を得ないで起動する。さらに、より望ましくは、読み出されたデータ等のアクセス結果を、アクセス先のPEのOSの介入を得ないで起動する。

【0033】さらに、本発明による他の望ましい態様では、アクセス元のPEのOSから発行されたI/O装置へのアクセス要求が、他のPE内のI/O装置へのアクセスを要求するとき、アクセス先のPE内のI/O装置におけるI/O動作を要求するコマンドをアクセス元のPEからアクセス先のPEに、これらのPEを結ぶネットワークを介して転送する。アクセス先のPEにおいては、上記コマンドにตอบสนองして、そのPE内のI/O制御装置に指示するためのI/O動作指定情報を、そのPE内のOSを介さないで直接生成し、生成されたI/O動作指定情報を用いてこのI/O制御装置を起動する。

【0034】さらに、本発明によるマルチプロセッサシステムでは、アクセス元のPEは、そのPE内のOSから発行された、アクセス先のPE内のI/O装置におけるI/O動作を要求するアクセス要求にตอบสนองして、そのI/O動作の実行を要求する少なくとも一つのコマンドを生成し、アクセス先のPEにネットワークを介して転送する回路有し、アクセス先のPEは、アクセス元のPEから上記コマンドが転送されたときに、アクセス先のPE内のI/O装置に上記コマンドが要求するI/O動作を実行させるためのI/O動作指定情報を、このコマンドに基づいて、かつ、このアクセス先のPEのOSの介入を得ないで生成し、生成されたI/O動作指定情報を用いて上記I/O制御装置を起動する回路を有する。

【0035】

【発明の実施の形態】以下、本発明に係るマルチプロセッサシステムを図面に示した実施の形態を参照してさらに詳細に説明する。

【0036】＜発明の実施の形態＞

(1) 装置の概要

図1aは本発明に係るプロセッサ間通信方法を適用するためのマルチプロセッサシステムの概略ブロック図である。本システムは、複数のプロセッシングエレメント

(以下、PEと呼ぶことがある)、例えば100、200(これらはPE0、PE15と呼ぶことがある)が、ネットワーク900により接続され、各PEは同じ構造

を有する。すなわち、各PEは、CPU110及び主記憶130、I/Oインタフェース回路120、I/O装置121を持つ。各PE内の主記憶130は、このシステムに共通の主記憶の一部を構成し、そのPEで実行されるプログラムおよびデータを保持するもので、このシステムはいわゆる分散メモリ型の並列計算機システムである。I/O装置121は、本実施の形態ではディスク記憶装置などの記憶装置であるが他の入出力装置でもよい。なお、図にはI/O装置は一つしか図示していないが、I/Oインタフェース回路120には複数のI/O装置が接続されていてもよい。I/Oインタフェース回路120は、CPUでのI/O命令が実行されたときに、I/O装置121へのアクセスコマンドを発行する回路で、ここでは、バス121aは、SCSIバスであり、I/Oインタフェース回路120はSCSIインタフェース回路上に形成されたディスク制御装置である。102は、送信コマンド処理回路、101は受信コマンド処理回路、103は受信したコマンドにตอบสนองしてI/O装置121への起動を要求する回路であり、いずれも本実施の形態に特有の回路であり、各PEが他のPE内のI/O装置をアクセスするときに、当該他のPE内のCPUを介さないで直接I/Oインタフェース回路120を介してI/O装置121をアクセスするのに使用される。

【0037】(2) 入出力動作の概要

この並列計算機システムの複数のPE内のI/O装置121は、データベースを分散して保持し、各PE内の主記憶130は、このデータベースを利用するためのアプリケーションプログラムと、オペレーティングシステムおよびそれらが使用するデータを保持する。オペレーティングシステムには、アプリケーションプログラムからの入出力装置へのアクセス要求を処理するファイルシステムと、そのファイルシステムからの入出力要求を処理するデバイスドライバとが含まれる。

【0038】本実施の形態における入出力動作の概要は以下の通りである。

【0039】(a) アクセス元のPE内のデバイスドライバの起動

アクセス元のPE内のファイルシステムは、アプリケーションプログラムから発行された入出力要求にตอบสนองして、I/O命令(ファイルシステムレベルのI/O命令)を発行して、アクセス元PE内のデバイスドライバを起動する。

【0040】(b) アクセス元PE内のI/O装置のアクセス

アクセス元OS内のデバイスドライバは、起動されると、上記I/O命令が指定するファイルがアクセス元のI/O装置121内にあるか否かを判別し、そのファイルがアクセス元のI/O装置121内にあるときには、デバイスドライバでは、このI/O命令が指定する、そ



のファイルの物理的な記憶位置を用いて、I/Oインタフェース回路120を起動するI/O命令（デバイスドライバレベルのI/O命令）を発行し、この回路120がI/O装置121内のこの記憶位置をアクセスする。

【0041】（c）他のPE内のI/O装置のアクセス上記ファイルシステムレベルのI/O命令が指定するファイルがアクセス元のI/O装置121内にないときには、アクセス元のPEのデバイスドライバは、以下のようにして、他のアクセス先のPE内のI/O装置を、アクセス先のPE内のOSを介さないでアクセスするためのネットワーク起動命令を発行する。以下では、以上の処理の詳細を述べる。

【0042】（3）アクセス元のPEのファイルシステムの動作の詳細

（a1）アクセス元PE内のファイルシステムは、アプリケーションプログラムから発行された入出力要求にตอบสนองして、主記憶130上に、読み出しデータあるいは書き込みデータを格納するためのバッファ領域を確保する。

【0043】（a2）アプリケーションプログラムが指定した、ファイルの論理的な名前より、そのファイルの物理的な記憶位置を求める。

【0044】（a3）その物理的な記憶位置から、そのファイルがアクセス元のI/O装置121内にあるかを判別する。そのファイルがアクセス元のI/O装置121内にあるときには、さらに、ファイルのロック処理などのファイル管理を行う。

【0045】（a4）そのファイルがアクセス元のI/O装置121内にないときには、アクセス元のファイルシステムは、そのファイルを保持するI/O装置を有する他のPEのファイルシステムとの間で、プロセッサ間通信を用いて交信し、そのファイルの物理的な位置を教えるとともに、そのファイルのアクセス権を当該他のPE内のファイルシステムより確保する。

【0046】アクセス元のファイルシステムとアクセス先のファイルシステムの間のプロセッサ間通信は、アクセス元のデバイスドライバとアクセス先のデバイスドライバとを介して、メッセージパッシングにより実行される。このとき、アクセス元PEおよびアクセス先PE内の送信コマンド処理回路コマンド102と受信コマンド処理回路101、I/O起動回路103が使用される。この時の装置動作の詳細は後に説明する。

【0047】（a5）そのファイルをアクセスするためのI/O命令を発行し、アクセス元OS内のデバイスドライバを起動する。

【0048】すなわち、アクセス元のPE内のファイルシステムは、読み出しアクセスあるいは書き込みアクセスの動作を指定するI/O動作指定情報（以下では、I/Oコントロールレコード（ICR）と呼ぶ）を作成し、主記憶130上のあらかじめ定められたレコード格

納領域に格納する。その後、このレコード格納領域のアドレスを指定するI/O命令を発行し、アクセス元のOS内のデバイスドライバを起動する。

【0049】読み出しアクセスのためのI/Oコントロールレコードは、図10に示されるように、フィールド4001にコマンド、フィールド4002内に先に確保された読み出し用のバッファのアドレス、フィールド4003に、読み出すべきデータの長さ、フィールド4004に、アクセスすべきI/O装置の番号及びその装置内のアクセスすべきデータの記憶位置を含む。フィールド4001内のコマンドは、読み出すべきファイルが自PE内にあるか否かにより、読み出しコマンドRあるいはリモート読み出しコマンドRRとなる。

【0050】書き込みアクセスのためのI/Oコントロールレコードは、図11に示されるように、フィールド4001にコマンドを含み、フィールド4002は、先に確保された書き込み用のバッファ内に書き込まれた書き込みデータのアドレスを保持する。他のフィールドは、読み出しアクセスのためのI/Oコントロールレコードと同じ内容を含む。フィールド4001内のコマンドは、アクセスすべきファイルが自PE内にあるか否かにより、書き込みコマンドWあるいはリモート書き込みコマンドRWとなる。

【0051】読み出しアクセスのためのI/Oコントロールレコード（図10）および書き込みアクセスのためのI/Oコントロールレコード（図11）のいずれの場合にも、フィールド4004に保持されたI/O装置番号は、全てのPE内の複数のI/O装置に対してユニークに定められた番号であり、アクセス元PEのデバイスドライバは、このI/O装置番号から、その番号のI/O装置が属するPEを判別する。

【0052】（4）アクセス元PE内のI/O装置のアクセスの詳細

アクセスすべきファイルがアクセス元のPE内にあるときには、アクセス元PEは上記処理（b）を実行する。ここでは、以下の処理が実行される。

【0053】（b1）アクセス元PE内のデバイスドライバは、上記I/O命令にตอบสนองして、この命令が指定するI/OコントロールレコードのアドレスをCPUICRアドレスレジスタ125に書き込む命令を発行した後、I/Oインタフェース回路120を起動するデバイスレベルのI/O命令を発行する。CPUは、上記書き込み命令にตอบสนองして、このアドレスをレジスタ125にバス115を介して書き込み、さらに、上記デバイスレベルのI/O命令にตอบสนองして線110a、ORゲート122を介してI/Oインタフェース回路120を起動する。線110a上の起動信号は、さらに直接I/Oインタフェース回路120に供給され、I/Oインタフェース回路120の起動元を識別するのに使用される。

【0054】（b2）I/Oインタフェース回路120

は、線110a上の起動信号により起動されると、信号123sを通じてセクタ123を切り換えて、CPUICRアドレスレジスタ125からI/Oコントロールレコードのアドレスを読み出し、主記憶130からバス115を介してI/Oコントロールレコードをフェッチする。その後、このレコードが読み出しコマンドRを含むときには、I/O装置121からデータを読み出し、このレコードにより指定された、主記憶内130内の読み出しバッファに書き込む。さらに、I/Oの終了状態信号120aを終了書込回路126に出力する。このレコードが書き込みコマンドWを含むときには、主記憶内のバッファ領域から書き込みデータを読み出し、このレコードにより指定された、I/O装置121内の記憶位置に書き込む。さらに、I/Oの終了状態信号120aを終了書込回路126に出力する。

【0055】(b3) 終了書込回路126は、マイクロコンピュータにより構成され、図16のフローを有するプログラムを実行する。すなわち、実行したコマンドが、自PE内のI/O装置への読み出しコマンドRあるいは書き込みコマンドWのいずれかであるか否かを判定する(ステップ5401)する。今の場合には、実行したコマンドは読み出しコマンドRまたは書き込みコマンドWと仮定しているので、主記憶130上の、I/O終了状態を書き込むための固定のI/O終了状態領域に終了状態を書き込み(ステップ5402)、その後線127sを通じてマルチプレクサ127をCPU側に切り換え、完了信号126aをマルチプレクサ127、線127a、ORゲート111、線111aを通じてCPUにI/O終了割込として送出する(ステップ5403)。

【0056】(b4) この後は従来と同様に、ファイルシステムがこの割り込みに応答して、読み出されたデータを要求元のアプリケーションプログラムに引き渡す。

【0057】(5) 他PE内のI/O装置へのアクセスの詳細

アクセスすべきファイルがアクセス元のPE以外の他のPE内にあるときには、上記処理(c)は以下のようにして実行される。

【0058】(c1) ネットワーク起動命令の発行

アクセス元のPE内のデバイスドライバは、アクセス元PE内のファイルシステムが発行した上記I/O命令が指定するI/O動作の実行を要求するために、アクセス先PEに以下のようにして一組のネットワークコマンドを送信することを要求する。すなわち、各ネットワークコマンドを指定するトランスファーコントロールワード(以下、TCWとよぶ)を主記憶130上に作成した後、このTCWをCPU TCWアドレスレジスタ193に書き込む命令を発行する。その後、ネットワーク起動命令を発行し、そのTCWが指定するコマンドの送信を要求する。この動作を異なるネットワークコマンドに関して順次行う。

【0059】(c2) ネットワークコマンドの送信

アクセス元PE内のCPUは、各TCWに対する書き込み命令を実行すると、そのTCWのアドレスをCPU TCWアドレスレジスタ193(図1C)にバス115を介して書き込み、その後にネットワーク起動命令を実行すると、送信起動信号を、線110b、ORゲート127bを通じて送信コマンド処理回路102を起動する。すなわち、リモートコマンド組立送信回路190は、信号192sを通じてセクタ192を切り換え、CPU TCWアドレスレジスタ193からTCWアドレスを読み出し、このアドレスを使用して主記憶130からTCWを読み出し、そのTCWに従い、送信すべきコマンドパケットを作成し、ネットワーク900に送出する。こうして、一つのネットワークのコマンドの送信を終えると、同じ動作を他のTCWに関して行う。

【0060】(c3) ネットワークコマンドの処理

アクセス先のPEでは、受信コマンド処理回路101と、I/Oアクセス起動要求回路103がネットワーク900から転送された一組のネットワークコマンドを、アクセス先のPEのOSを介さないで直接処理し、アクセス先のPE内のI/Oインタフェース回路120を起動する。この一組のネットワークコマンドが、データの書き込みに対するものであるときには、これらの一組のネットワークコマンドの一つは、書き込みデータを含み、I/Oインタフェース回路120はこの一つのネットワークコマンドに含まれた書き込みデータをI/O装置121に書き込む。また、この一組のネットワークコマンドが、データの読み出しに対するものであるときには、I/Oインタフェース回路120はこの一組のネットワークコマンドにより要求されるデータをI/O装置121から読み出す。読み出したデータを含むネットワークコマンドをネットワーク900を介してアクセス元PEに転送する。

【0061】(c4) 返信用コマンドの送信

アクセス先のPEでの上記コマンドの実行結果をアクセス元のPEに通知するコマンドをパケットの形で返送する。

【0062】(c5) 返信用コマンドの処理

アクセス元のPEでは返送されたコマンドを実行する。

【0063】(5A) 他のPE内のI/O装置へのデータの書き込み動作の詳細

アクセス元のファイルシステムが発行したI/O命令がデータの書き込み命令である時には、上記処理(c)は以下のように実行される。この場合には、アクセス元のデバイスドライバは、データ書き込みコマンドWDおよびI/O装置書き込みコマンドIOWを送ることを要求する。

【0064】(c1A1) データ書き込みコマンドWD用TCWの作成

先ず、データ書き込みコマンドWDを送信するためのT

CWを主記憶130上に作成する。このTCWは、図4に示すフォーマットを有し、宛先フィールド2001はアクセスすべきI/O装置を持つPEの番号を示す。フィールド2002は、データ書き込みコマンドWDを含み、データアドレス2004は、主記憶130上の書込データの先頭アドレスを指し、フィールド20052205は、そのデータのデータ長を指す。

【0065】(c2A1) データ書き込みコマンドWD用パケットの送信

リモートコマンド組立送信回路190は、このTCWが作成された後にデバイスドライバから発行されたネットワーク起動命令に応答して、このTCWのアドレスを先に述べた方法で読み出し、データ書込コマンドWD用のパケット(図5)を作成し、ネットワーク900を介して宛先フィールド3001のPEに送信する。ここで、フィールド3002は、送り元PEの番号を含み、この番号はレジスタ191から与えられる。フィールド3003はコマンドWDを含み、む。ここで、3001~3006をパケットヘッダと呼ぶ。フィールド3007から3008は転送すべき書き込みデータを含む。このデータは、リモートコマンド組立送信回路190がTCWに従い主記憶130から読み出す。その他のフィールドは図4のものと同様である。

【0066】(c3A1) 受信データの主記憶への格納  
アクセス先PE内の受信コマンド処理回路101では、リモートコマンドパケット受信分解回路170(図1b)がこのデータ書き込みコマンドWD用のパケットを分解し、データ書き込みコマンドWDを線170dにおよびパケットのヘッダ情報を線170hに、書き込みデータを線170gにそれぞれ出力する。受信コマンド処理回路101では、バッファアドレスポインタ185は、受信した書き込みデータを格納するための、主記憶130上にあらかじめ定められたバッファ領域内の空き領域の先頭アドレスを保持する。このアドレスポインタの内容は、リモートデータ書き込み回路180および後に説明するICR作成回路140により更新されるようになっている。なお、このバッファ領域は以下の述べる読み出しデータその他の、リモートからのI/Oアクセス関連するデータを主記憶に保持するのにも使用される。なお、バッファ領域は、PE間の通信のためにあらかじめ主記憶上に確保されるバッファ領域とは別のバッファ領域を使用することが望ましい。

【0067】リモートデータ書込回路180はマイクロコンピュータにより構成され、リモートコマンドパケット受信分解回路170から受信したコマンドが入力されるごとに、図12に示すフローを有するプログラムを実行する。すなわち、受信したコマンドがデータ読み出しコマンドRDか否かを判定する(ステップ5001)。今の場合、受信したコマンドはデータ書き込みコマンドWDであるため、バッファアドレスポインタ185内の

ポインタを線185aを介して読み出し、主記憶130内にこのアドレスから始まるバッファ領域をアロケートした後(ステップ5004)、受信した書き込みデータ(170g)をこのバッファに書き込む(ステップ5005)。その後、線181sを通じてマルチプレクサ181を切り換え、ステップ5004でアロケートしたバッファのアドレスを線180aを介してWDデータアドレスレジスタ182に書き込む(ステップ5007)。さらに、線185aを介して、バッファアドレスポインタ185を受信したデータの長さだけ更新する。以上の処理により、データ書き込みコマンドWD用のパケットの送信及び処理が終了する。

【0068】(c1A2) I/O装置書き込みコマンドIOW用TCWの作成

アクセス元のデバイスドライバは、次に、I/O装置書き込みコマンドIOWを送信するためのTCWを主記憶130上に作成する。このTCWは、図6に示すフォーマットを有し、フィールド2002は、コマンドIOWを含み、フィールド2004は、上記I/O命令が指定した前記のI/OコントロールレコードICRのアドレスを含み、フィールド2005はこのレコードの長さを含む。アクセス元のデバイスドライバは、このTCWに対する前述の書き込み命令とネットワーク起動命令を発行する。

【0069】(c2A2) I/O装置書き込みコマンドIOW用パケットの送信

アクセス元のCPUは、この書き込み命令とネットワーク起動命令を前述した方法で実行して、このTCWのアドレスをCPU TCWアドレスレジスタ193に書き込み、リモートコマンド組立送信回路190をトリガする。リモートコマンド組立送信回路190は、データ書き込みコマンドの場合と同様にして、図7に示すフォーマットの、I/O装置書き込みコマンドIOW用のパケットを作成し、ネットワーク900に送信する。このパケットのフィールド3007から3008は、上記TCWが指定したI/OコントロールレコードICRを含む。

【0070】(c3A2) I/O装置へのアクセスの実行

アクセス先PE内のリモートコマンドパケット受信分解回路170は、このパケットを分解し、IOWコマンド解読信号を線170bに、パケットのヘッダ情報を線170hに、I/OコントロールレコードICRを線170gに出力する。

【0071】ICR作成回路140は、マイクロコンピュータにより構成され、リモートコマンドパケット受信分解回路170からI/O装置書き込みコマンドIOWあるいは後述するI/O装置読み出しコマンドIORの解読信号が線170bあるいは170cから入力されるごとに、図14に示すフローを有するプログラムを実行

し、I/Oインタフェース回路120に与えるべき新たなI/OコントロールレコードICRを作成し、このI/Oインタフェース回路120を起動する。この新たなI/OコントロールレコードICRは、リモートコマンドパケット受信分解回路170から線170gを介して与えられる受信したI/OコントロールレコードICRを使用して作成され、受信したI/OコントロールレコードICRがデータ読み出し用(図11)かあるいはデータ書き込み用(図12)かにより、図11あるいは図12のフォーマットを有する。今の例では、図11のフォーマットのI/OコントロールレコードICRを作成する。

【0072】すなわち、まず、バッファアドレスポインタ185より、作成されるI/OコントロールレコードICRを格納するための領域をアロケートした後(ステップ5201)、受信したコマンドがI/O装置書き込みコマンドIOWか否かを判定(ステップ5202)する。今の場合には、受信したコマンドがI/O装置書き込みコマンドIOWであるので、ステップ5203において以下のようにしてI/OコントロールレコードICRを作成する。まず、このレコードのフィールド4001には、リモートデータ書き込みコマンドRWを格納し、書込バッファアドレスフィールド4002には、先に送信されたデータ書き込みコマンドWDを受信したときに、WDデータアドレスレジスタ182に記憶されたアドレスを使用する。このアドレスは、受信した書込データが書き込まれた主記憶内のバッファ領域のアドレスである。従って、受信したパケット(図7)内のフィールド3007から3008に含まれていたI/OコントロールレコードICR(図11)内の書き込みアドレスバッファ4002は使用されることが分かる。さらに、ステップ5203において作成されるI/OコントロールレコードICRのフィールド4003、4004には、受信したI/Oコントロールレコード内のデータ長およびI/O装置番号、アクセスデータの物理的な記憶位置をそのまま使用する。この様に、作成されたI/Oコントロールレコードは、アクセス先のPE内のファイルシステムがデータ書き込みのためのI/O命令を発行する前に生成するI/Oコントロールレコードと同じデータ構造を有する。作成された新たなI/Oコントロールレコードを主記憶装置130に格納し、リモートICRアドレスレジスタ124(図1A)に線140bを介してその格納アドレスをセットした後(ステップ5204)、完了信号140aを出力し(ステップ5205)、ORゲート122を介してI/Oインタフェース回路120を起動する。完了信号140aは、さらに直接I/Oインタフェース回路120に供給され、I/Oインタフェース回路120の起動元を識別するのに使用される。

【0073】I/Oインタフェース回路120は、完了

信号140aにより起動されると、信号123sを通じてセクタ123を切り換え、リモートICRアドレスレジスタ124に保持されたアドレスを使用した、作成されたI/OコントロールレコードICRを主記憶130から読み出す。その読み出されたレコードにより指定された書込データバッファアドレスを使用して、先に受信した書き込みデータを主記憶130から読み出して、I/O装置121へ書き込む。I/O装置121へのアクセスが終了すると、I/O終了状態信号120aを出力する。

【0074】(c4A)ステータスコマンドストアコマンド用パケットの返送

返信用TCW作成回路150は、マイクロコンピュータにより構成され、先にリモートコマンドパケット受信分解回路170により線170bあるいは170cにI/O装置書き込みコマンドIOWの解読信号あるいはI/O装置読み出しコマンドIORの解読信号が出力されたときに、それらのコマンドの実行と並行して、図15に示すフローを有するプログラムを実行し、それらのコマンドの実行後に送信されるべき返信用パケットの作成に使用される返信用TCWを作成する。すなわち、図15に示すフローにおいて、まず、バッファアドレスポインタ185を用いて返信用TCWを格納するための領域をアロケートした後(ステップ5301)、受信したコマンドがI/O装置書き込みコマンドIOWか否かを判別する(ステップ5302)。今の場合には、受信したコマンドがI/O装置書き込みコマンドIOWであるので、図8で示すステータスコマンドストアコマンドSTを送付するためのTCWを作成する(ステップ5303)。すなわち、TCWの宛先領域2001には、信号170hを通じて送られてきたI/O装置書き込みコマンドIOWの送り元PE番号を格納し、フィールド2002には、このアクセス先PEの番号を格納し、フィールド2003には、ステータスコマンドストアコマンドSTを格納し、フィールド2005にはデータ長として値0を格納する。その後、この作成したTCWを主記憶130に格納し、その格納アドレスを返信用TCWレジスタ194にセットする(ステップ5304)。

【0075】終了書込回路126は、マイクロコンピュータにより構成されI/O装置121へのアクセスが終了したことに伴いI/Oインタフェース回路120がI/O終了状態信号120aを出力したときに起動され、図16に示されたフローを有するプログラムを実行する。すなわち、ステップ5401において、今回実行されたコマンドが、読み出しコマンドあるいは書き込みコマンドWではなく、リモート書き込みコマンドであると判断すると、信号126bを通じて、実行されたコマンドが正常に終了したか否かを表す終了状態を終了状態レジスタ195に書き込む(ステップ5404)。さらに、信号127sを通じてマルチプレクサ127を送信

コマンド処理回路102側に切り換えると同時に完了信号126aを出力し、線127b、ORゲート196、線196aを通じて送信コマンド処理回路102内のリモートコマンド組立送信回路190をトリガする(ステップ5405)。線127a上の起動信号は、また、起動元を識別するための信号として直接リモートコマンド組立送信回路190に供給される。

【0076】リモートコマンド組立送信回路190は、線127b上の起動信号により起動されると、信号192sを通じてセクタ192を切り換え、返信用TCWアドレスレジスタ194内のアドレスを読み出し、このアドレスを使用して主記憶130から、先に作成された返信用TCWを読み出す。さらに、終了状態レジスタ195から終了状態を読み出し、このTCWと終了状態とから、返送すべきステータスコマンドストアコマンドST用のパケット(図9)を完成し、アクセス元のPEに送出する。

【0077】(c5A)ステータスコマンドストアコマンドの実行  
アクセス元のPEでは、リモートコマンドパケット受信分解回路170により、このコマンドST用のパケットを分解し、STコマンド解読信号を線170fに、パケットのヘッダ情報を線170hに出力する。

【0078】状態書込回路160は、マイクロコンピュータにより構成され、ステータスコマンドストアコマンドSTの解読信号あるいは後に述べるリードコマンドRDの解読信号が線170fあるいは線170eから供給される度に、図13に示すフローを有するプログラムを実行する。すなわち、線170fに出力された、パケットのフィールド3006内の終了状態を主記憶130上のI/O終了状態領域に書き込み(ステップ5101)、受信したコマンドがステータスコマンドストアコマンドSTか否かを判定し(ステップ5102)、今の場合のように受信したコマンドがステータスコマンドストアコマンドSTの場合、完了信号160aを出力し、ORゲート111a経由で、CPUにI/O終了割り込みをかける(ステップ5103)。その後は、この割り込みに応答して、アクセス元のファイルシステムが、従来と同様に、上記I/O終了状態領域に書き込まれた終了状態を読み出し、先に発行したI/O命令の実行が完了したか否かを判別する。

【0079】以上の処理により、アクセス元のPEは、リモートPEの持つI/O装置へ、そのPEのOSの介入を得ないで、直接データを書き込むことができる。

【0080】(5B)他のPE内のI/O装置からのデータの読み出し動作の詳細

次に、他PEの持つI/O装置への読出アクセスのときに実行される上記処理(c)を、上記書き込みアクセス時ととの相違点を中心に説明する。

【0081】(c1B)I/O装置読み出しコマンドI

OR用TCWの作成

この読み出し動作のために、アクセス元のPE内のファイルシステムが上記処理(a)で作成するI/Oコントロールレコードは、図10に示すフォーマットを有し、フィールド4001内にリモート読み出しコマンドRRを含む。アクセス元のPE内のデバイスドライバは、作成上記処理(c1)において、このI/Oコントロールレコードを使用して、図6に示すI/O装置読み出しコマンドIOR用のTCWを作成する。このTCWは、I/O装置書き込みコマンドIOW用のTCWと同じフォーマットであり、そのTCWと同様に作成される。

【0082】(c2B)I/O装置読み出しコマンドIOR用パケットの送信

その後、アクセス元のCPUはネットワーク900にこのコマンド用のパケットを送出する。このパケットは、図7に示すように、I/O装置書き込みコマンドIOW用のパケットと同じフォーマットを有し、パケットの送出手順もそのパケットの場合と全く同じである。

【0083】(c3B)I/O装置読み出しコマンドIORの実行

アクセス先PEでは、リモートコマンドパケット受信分解回路170により、このパケットを分解し、I/O装置読み出しコマンドIORの解読信号を線170cに、パケットのヘッダ情報を線170hに、I/OコントロールレコードICRを線170gを出力する。ICR作成回路140は、図14に示すフローに基づき、送付されたI/OコントロールレコードICRを使用してI/Oインタフェース回路120を起動する。

【0084】すなわち、まず、バッファアドレスポインタ185より、ICRを受信するための領域をアロケートした後(ステップ5201)、受信したコマンドがI/O装置書き込みコマンドIOWでないと判定する(ステップ5202)と、バッファアドレスポインタ185より、I/OコントロールレコードICRのデータ長フィールド(図10、4003)の値の大きさのバッファをアロケートする(ステップ5206)。このバッファは、I/O装置121から読み出されたデータを一時的に格納するために用いられる。その後、このバッファのアドレスを示すRDアドレス信号を線140cを介して返信用TCW作成回路150に出力する(ステップ5207)。このRDアドレス信号は、後で説明する返信用TCWを作成するために用いられる。さらに、ステップ5208において、受信したコマンドを実行するために、図10に示すフォーマットを有する、リモート読み出しコマンドRR用の新たなI/OコントロールレコードICRを作成する。すなわち、上記バッファのアドレスを、この新たなレコードの読出バッファアドレスフィールド4002として使用し、送付されたI/OコントロールレコードICRのデータ長(4003)、I/O装置番号及びアクセスデータの物理的な記憶位置(40

04)も使用する(ステップ5208)。この新たなI/OコントロールレコードICRは、アクセス先のPE内のファイルシステムがデータ読み出しのためのI/O命令を発行する前に作成するものと同じ構造を有する。その後リモートICRアドレスレジスタ124に作成したレコードのアドレスをセットした後(ステップ5204)、完了信号140aを出力し(ステップ5205)、I/Oインタフェース回路120を起動する。

【0085】I/Oインタフェース回路120は、I/O装置書き込みコマンドIOWの場合と同様にして、I/O装置への読出アクセスを行い、I/O装置から読み出されたデータを、主記憶130に保持された、先に作成されたI/OコントロールワードICR内の読出バッファアドレスフィールド4002を使用して主記憶130上のバッファ領域に書き込む。その後の動作はI/O装置書き込みコマンドIOWの場合と同様である。

【0086】(c4B)リードコマンド用パケットの返送

返信用TCW作成回路150では、図15に示すフローに基づいて、まず、バッファアドレスポインタ185を用いて返信用TCWを格納するための領域をアロケートした後(ステップ5301)、実行されたコマンドが、I/O装置書き込みコマンドIOWでないことを判定する(ステップ5302)と、図2で示すデータ読み出しコマンドRDを送付するためのTCWを作成する(ステップ5305)。すなわち、リモート書込アドレス2003には、このコマンドにより送られてきた後に、信号170gを通じてリモートコマンド受信分解回路170より伝達された、受信したI/Oコントロールレコード(図10)内の読出バッファアドレス4002が含まれる。このアドレスは、アクセス元PEにおいて、I/O装置から読み出されたデータを格納するための主記憶内アドレスである。さらにデータアドレス2004は、RDアドレス信号140cで送られてきたI/O装置121から先に読み出されたデータを格納したバッファのアドレスが格納される。その他のフィールドは、実行されたコマンドが、I/O装置書き込みコマンドIOWの場合に作成されるTCW(図4)と同じである。その後、返信用TCWレジスタ194に作成したTCWのアドレスをセットする(ステップ5304)。

【0087】リモートコマンド組立送信回路190は、I/O装置書き込みコマンドIOWの場合と同様にして、I/Oアクセスの終了により送信が起動されると、I/O装置書き込みコマンドIOWの場合と同様にして、データ読み出しコマンドRD用のパケット(図3)を作成し、アクセス元のPEに送出する。但し、I/O装置書き込みコマンドIOWの場合と異なり、このパケットには、I/O装置121から主記憶130に読み出されたデータが含まれる。

【0088】(c5B)データ読み出しストアコマンド

の実行

アクセス元のPEでは、リモートコマンドパケット受信分解回路170により、このパケットを分解し、データ読み出しコマンドRDの解読信号を線170eに出力し、パケット内の読み出しデータを線170gに出力する。パケットのヘッダ情報を線170hに出力する。RD信号を受け取った状態書込回路160では、図13に示すフローに基づき、主記憶130上のI/O終了状態領域にパケットの状態領域3006の値を書き込む(ステップ5101)。さらに、受信したコマンドがステータスコマンドストアコマンドSTでないことを判定し(ステップ5102)、処理を終了する。

【0089】リモートデータ書込回路180では、RDコマンド解読信号に応答して、図12に示すフローに基づき、受信したコマンドがデータ読み出しコマンドRDであることを判定し(ステップ5001)、受信したパケットのリモート書込アドレスフィールド3004の示す主記憶130上の領域に、受信したパケットのデータフィールド内の読み出しデータを書き込んだ後(ステップ5002)、完了信号180bを出力し、信号111a経由でCPUにI/O終了割込をかける。その後は、この割込みに応答して、アクセス元のファイルシステムが、従来と同様に、上記I/O終了状態領域に書き込まれた終了状態を読み出し、先に発行したI/O命令の実行が完了したか否かを判別する。完了した場合には、読み出されたデータを要求もとのアプリケーションプログラムに引き渡す。

【0090】以上の処理により、アクセス元のPEのファイルシステムは、リモートPEの持つI/O装置から、そのPEのOSの介入を得ないで自PEのバッファ領域に直接データを読み出すことができる。

【0091】(6)メッセージパッシング  
上記処理(a)で、プロセッサ間通信に使用される、メッセージパッシング用のコマンドMPとその処理について説明する。まず、アクセス元のファイルシステムは、主記憶130上に他のPEに転送したいデータを主記憶130上に作成した後に、コマンドMPのためのTCW(図4)を作成した後、TCWのアドレスをCPU TCWアドレスレジスタ193に書き込む命令を発行する。ここで、データアドレス2004は転送するデータの先頭アドレスである。さらに、ネットワーク起動命令を発行する。CPUは上記書き込み命令を実行してレジスタ193にTCWのアドレスを書き込み、ネットワーク起動命令を実行したときに、送信起動信号を線110b、ORゲート196、線196aを通じて送信コマンド処理回路102内を通じリモートコマンド組立送信回路190をトリガする。線110b上の起動信号は、また、起動元を識別するための信号として直接リモートコマンド組立送信回路190に供給される。その後は、前述のデータ書き込みコマンドWDの場合と同様にして、

このコマンドを含む図5のパケットを送信先PEに転送する。

【0092】送信先PEでは、このパケットを受信すると、リモートデータ書き込み回路180では、図12に示すフローに基づき、前述のデータ書き込みコマンドWDの場合と同様にして、受信したデータを主記憶130に書き込む(ステップ5001から5005)。その後は、前述のデータ書き込みコマンドWDの場合とは異なり、信号181sを通じてセクタ181を切り換え、ステップ5004でアロケートしたバッファのアドレスを信号180aを介してMP受信データアドレスレジスタ183に書込んだ後(ステップ5008)、完了信号180bを介してCPUにメッセージ受信割込をかける(ステップ5009)。

【0093】割込を受け取ったCPUは、MP受信データアドレスレジスタ183を読み出すことにより、メッセージを受信したバッファアドレスを知り、受信メッセージの内容を読み出すことができる。

【0094】以上述べたように、本実施の形態では、他のPEの持つI/O装置への、書込、読出アクセスを他のPE内のOSを介さないで行うことが可能になる。これにより、他のPEのOSの介入を得ないで、当該他のPEのI/O装置へのアクセスを行うことが可能になり、各PE内のCPUの実行効率を大幅に向上させることができる。

#### 【0095】<変形例>

本発明は以上の実施の形態に限定されるのではなくいろいろの変形例にも適用可能である。例えば、

(1) 以上においては、他のPE内のI/O装置へのデータの書き込みを要求するI/O命令が実行されたときには、書き込みデータコマンドとI/O装置書き込みコマンドとをアクセス先のPEに送信した。しかし、これらのコマンドの内容を一つのコマンドに含ませて送ることも可能である。

【0096】(2) また、以上においては、OSのファイルシステムがファイルの物理的な位置を求めたり、ファイルのロック処理などのファイル管理を行う。しかし、アプリケーションプログラムが直接ファイルを管理し、デバイスドライバがアプリケーションプログラムから直接呼ばれる計算機システムもある。このようなI/O動作はraw I/Oと呼ばれる。本発明はそのようなシステムにも適用可能である。

【0097】(3) 上記実施の形態において、計算機間ネットワーク等の通信機構への入出力インタフェース回路をこのI/Oインタフェース回路120に付加してあるいはその代わりに使用するシステムにも本発明を適用することができる。

【0098】(4) 上記実施の形態で使用したリモートデータ書き込み回路等のいくつかの回路は、マイクロコンピュータで形成されているが、これらを専用の論理回

路により構成することも可能である。

【0099】(5) 上記の実施の形態では、PE内のCPUの数は1個であるが、PE内が複数のCPUを持つ主記憶共有型のマルチプロセッサで構成されていても良い。

【0100】(6) 上記の実施の形態では、I/O装置をアクセスするのは、該当するマルチプロセッサシステム内の他のPEのCPUであったが、計算機間のネットワークにつながった他の計算機からのアクセス要求に関しても、本発明は適用可能である。

【0101】ここで、該当する外部の計算機が接続されているのは、I/O装置を持つPEそのものでも、他のPEでも良い。

【0102】(7) (6)において他のPEに接続された計算機間ネットワークの先の計算機からアクセスされる場合、内部ネットワーク900を通じてI/O装置を持つPEにアクセス要求を転送する必要があるが、その際の内部ネットワークと外部ネットワークの乗り換えにも本発明の技術を使用することができる。

#### 【0103】

【発明の効果】本発明によれば、複数のプロセッシングエレメントからなる計算機システムにおいて、他のPEの持つI/O装置にアクセスする際に、他PEのI/Oインタフェース回路に対してI/O装置を持つPEのCPUの助けをかりずにハードウェアで直接起動できる。

#### 【図面の簡単な説明】

【図1a】本発明のI/Oアクセス方式を実現する計算機システムの概略ブロック図である。

【図1b】図1aのシステムに使用する受信コマンド処理回路の概略ブロック図である。

【図1c】図1aのシステムに使用する送信コマンド処理回路の概略ブロック図である。

【図2】ネットワークへRDコマンドを出すためのTCWのフォーマットである。

【図3】ネットワーク上のRDコマンドのパケットフォーマットである。

【図4】ネットワークへMPもしくはWDコマンドを出すためのTCWのフォーマットである。

【図5】ネットワーク上のMPもしくはWDコマンドのパケットフォーマットである。

【図6】ネットワークへIORもしくはIOWコマンドを出すためのTCWのフォーマットである。

【図7】ネットワーク上のIORもしくはIOWコマンドのパケットフォーマットである。

【図8】ネットワークへSTコマンドを出すためのTCWのフォーマットである。

【図9】ネットワーク上のSTコマンドのパケットフォーマットである。

【図10】I/Oインタフェース回路にRもしくはRRコマンドを出すためのICRのフォーマットである。

チャートである。

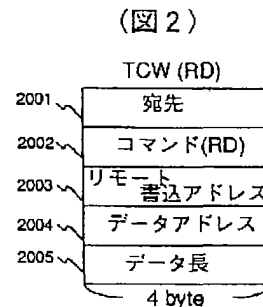
【図 15】返信用TCW作成回路150の動作の動作のフローチャートである。

【図 16】終了書込回路 126 の動作の動作のフローチャートである。

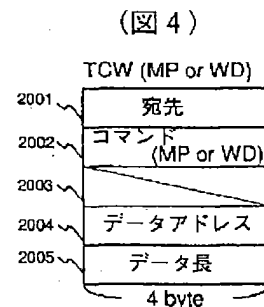
・【符号の説明】

100、200……プロセッシングエレメント (PE)

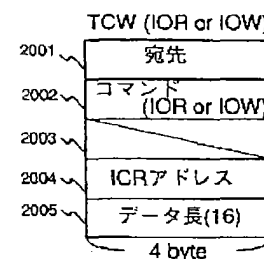
【図 2】



【図 4】

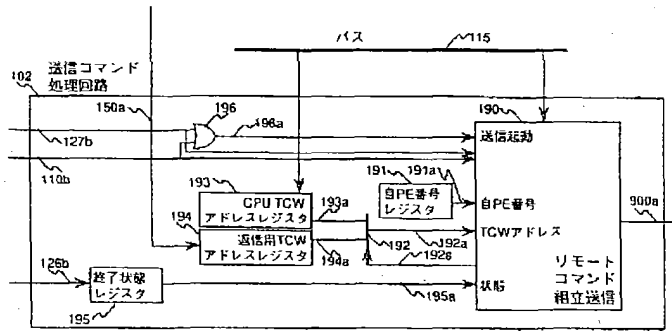


(圖 6)



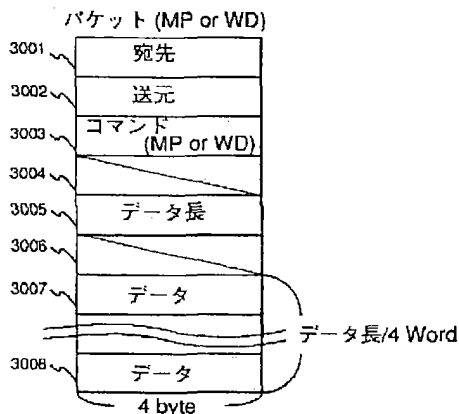


【図1c】



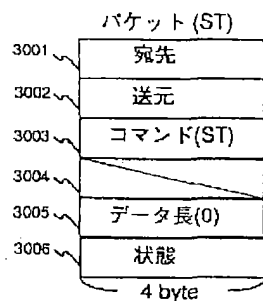
【図5】

(図5)



【図9】

(図9)

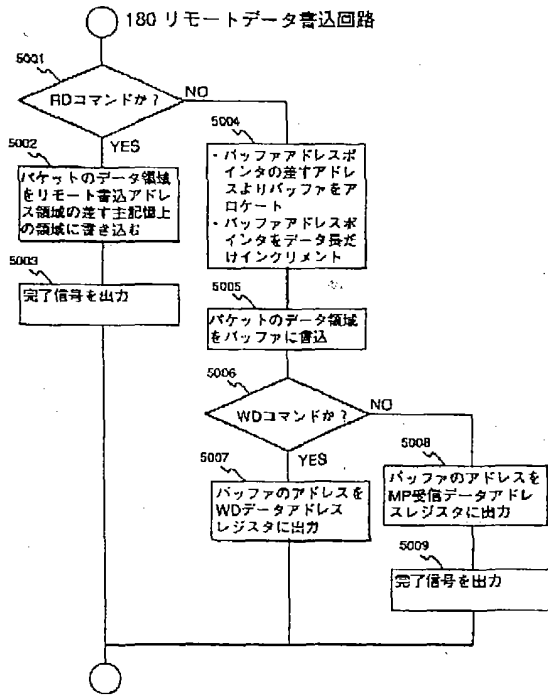


【図10】

(図10)

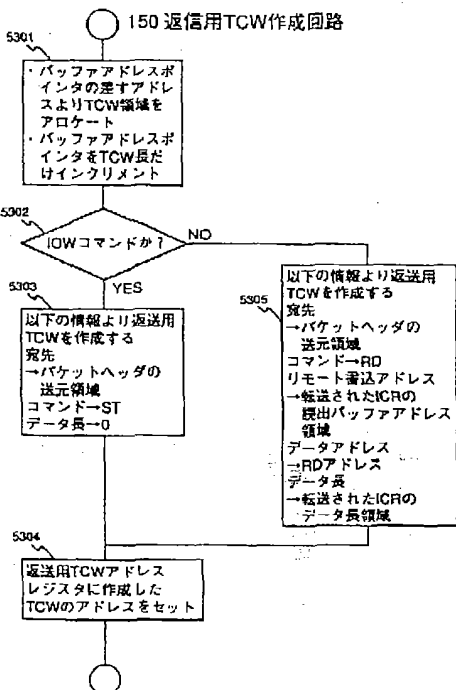
【図12】

(図12)



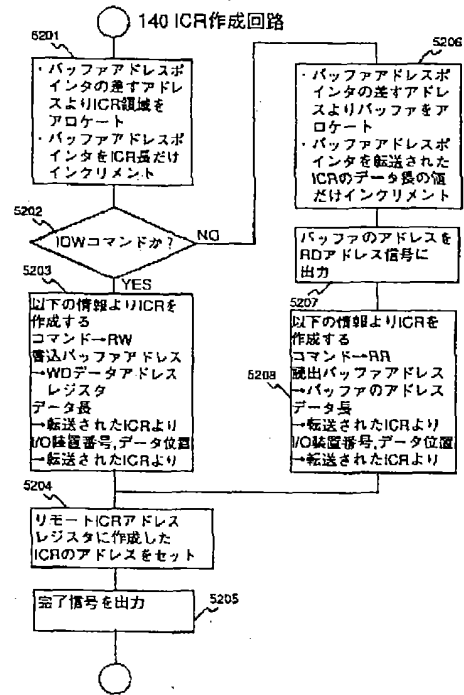
【図15】

(図15)



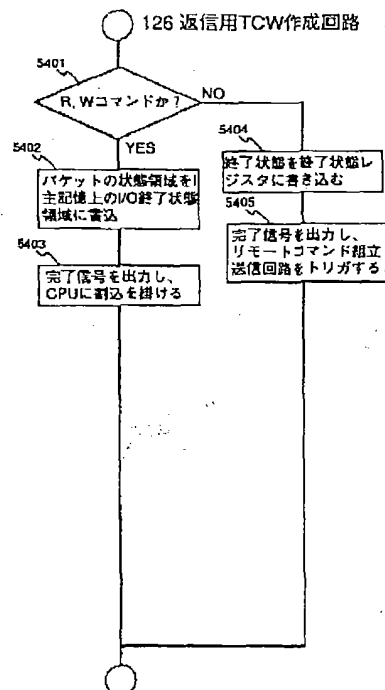
【図14】

(図14)



【図16】

(図16)



フロントページの続き

(72)発明者 樋口 達雄

東京都国分寺市東恋ヶ窪一丁目280番地  
株式会社日立製作所中央研究所内

(72)発明者 村橋 英樹

東京都国分寺市東恋ヶ窪一丁目280番地  
株式会社日立製作所中央研究所内